

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 10-097286
 (43)Date of publication of application : 14.04.1998

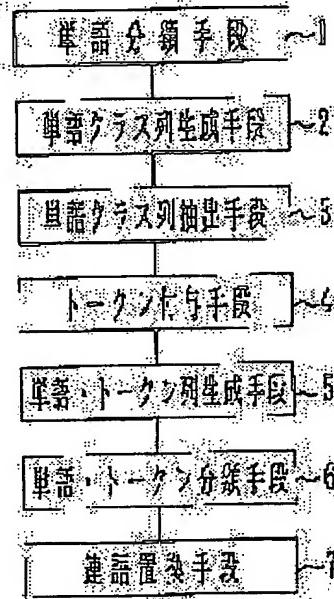
(51)Int.CI. G10L 3/00
 G10L 3/00
 G06F 17/28

(21)Application number : 09-167243 (71)Applicant : FUJITSU LTD
 (22)Date of filing : 24.06.1997 (72)Inventor : SHIODA AKIRA

(30)Priority
 Priority number : 08204986 Priority date : 02.08.1996 Priority country : JP

(54) WORD AND COMPOUND WORD CLASSIFYING PROCESSING METHOD, COMPOUND WORD EXTRACTING METHOD, WORD AND COMPOUND WORD CLASSIFYING PROCESSOR, SPEECH RECOGNITION SYSTEM, MACHINE TRANSLATING DEVICE, COMPOUND WORD EXTRACTING DEVICE, AND WORD AND COMPOUND WORD STORAGE MEDIUM

(57)Abstract:
PROBLEM TO BE SOLVED: To make speech recognition and machine translation accurate by classifying words and compound words included in text together and generating a class wherein the words and compound word are mixed.
SOLUTION: The word and compound word classifying processor consists of a word classifying means 1, a word class string generating means 2, a word class string extracting means 3, a token giving means 4, a word and token string generating means 5, a word and token classifying means 6, and a compound word substituting means 7. Word classes obtained by classifying words are mapped in a linear array of words of the text data to generate a linear array of word classes. In the linear array of the word classes of the text data, word class arrays which all have adherence above a specific value between adjacent word classes are extracted and tokens are given to the word class arrays. The words and tokens are classified together and then a word class array corresponding to a token is substituted by a couple belonging to the word string. Namely, a classifying process can be performed automatically without discriminating between words and compound words.



LEGAL STATUS

- [Date of request for examination] 15.07.2003
- [Date of sending the examiner's decision of rejection]
- [Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
- [Date of final disposal for application]
- [Patent number]
- [Date of registration]
- [Number of appeal against examiner's decision of rejection]
- [Date of requesting appeal against examiner's decision of rejection]
- [Date of extinction of right]

(19)【発行国】日本国特許庁 (JP)

(12)【公報種別】公開特許公報 (A)

(11)【公開番号】特開平10-97286

(43)【公開日】平成10年(1998)4月14日

(54)【発明の名称】単語・連語分類処理方法、単語・連語抽出方法、単語・連語分類処理装置、音声認識装置、機械翻訳装置、連語抽出装置及び単語・連語記憶媒体

(51)【国際特許分類第6版】

G10L 3/00 561 5621 G06F 17/28.

[FI]

G10L 3/00 561 G 521 C G06F 17/38

[審査請求未請求]

(請求項の数) 17

[出願形態] OI

[全頁数] 24

(21)[出願番号]特願平9-167243

(22)[出願日]平成9年(1997)6月24日

(31)[优先権主張番号]特願平8-204986

(32)[优先日]平8(1996)8月2日

(33)[优先権主張国]日本(JP)

(71)[出願人]

[識別番号]0000005223

[氏名又は名称]富士通株式会社
[住所又は居所]神奈川県川崎市中原区上小田中4丁目1番1号

(72)[発明者]

[氏名]鶴田 明

[住所又は居所]神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内

(74)[代理人]

[弁理士]

[氏名又は名称]大曾 義之 (外1名)
[出願人]

の単語とトークンとの一次元列において、互いに異なる単語と互いに異なるトークンとを抽出し、前記単語と前記トークンが混在する集合を単語・トークンクラスに分割した第2のクラスリラングを生成するステップと、前記テキストデータに存在する単語列のうち、前記トークンに對応するものを連語として抽出し、前記単語・トークンクラスの中のトークンを前記連語で置換することにより、前記単語と前記連語が混在する集合を単語・連語クラスに分割した第3のクラスリラングを生成するステップとを備えることを特徴とする。

【請求項2】前記第1のクラスリラングは、前記単語クラスの平均相互情報量に基づいて生成されることを特徴とする。

【請求項3】前記第2のクラスリラングは、前記単語・トークンクラスの平均相互情報量に基づいて生成されることを特徴とする請求項1に記載の単語・連語分類処理方法。

【請求項4】テキストデータに含まれる単語を分類した単語クラスを前記単語クラスと、前記テキストデータの単語の一次元列にマッピングして単語クラスの一次元列を生成するステップと、前記テキストデータの単語クラスの一次元列において、隣接する単語クラス間の粘着度が全て所定値以上の単語クラス列を、前記テキストデータの単語クラスの一次元列から抽出するステップと、前記テキストデータに含まれる単語と前記単語クラス列とを構成する個々の単語クラスから、前記テキストデータに隣接して存在する個々の単語クラス列を前記単語クラス列に属する連語で置換するステップと、前記単語クラス列を前記連語クラス列に属する連語で置換するステップとを備えることを特徴とする。

【請求項5】テキストデータに含まれる単語を分類した単語クラスを前記単語クラスと、前記テキストデータの単語の一次元列にマッピングして単語クラスの一次元列を生成するステップと、前記テキストデータの単語クラスの一次元列において、隣接する単語クラス間の粘着度が全て所定値以上の単語クラスの一次元列から抽出する単語クラス列を、前記テキストデータの単語クラスの一次元列から抽出するステップと、前記単語クラスから、前記テキストデータに隣接して存在する個々の単語別々に取り出して連語を抽出するステップとを備えることを特徴とする。

【請求項6】テキストデータの単語から互いに異なる単語を抽出し、抽出された前記単語を分割して単語クラスを生成する単語分類手段と、前記単語クラスが属する前記単語クラスで置換することにより、前記テキストデータの単語の一次元列を生成するステップと、前記テキストデータの単語の一次元列において、隣接する単語クラス間の粘着度が全て所定値以上に達する単語クラスを、前記テキストデータの単語の一次元列から抽出手段と、前記単語クラスの単語の一次元列にマッピングして単語クラス間の粘着度が全て所定値以上の単語クラス列を、前記テキストデータの単語の一次元列から抽出された各単語クラス列にトークンを付与するトークン付与手段と、前記テキストデータの単語の一次元列のうち、前記単語クラス列に属する単語クラスを前記トークンで置換することにより、前記単語クラス列の単語・トークンの一次元列を生成する単語・トークンクラスと、前記テキストデータの単語・トークンの一次元列に含まれる単語とトークンとが混在する集合を分割して単語・トークンクラスを生成する単語・トークン分類手段と、前記単語・トークンクラスの中のトークンを、前記単語・トークン生成手段により置換された単語列に逆置換して連語を生成することを特徴とする。

【請求項7】前記単語分類手段は、前記テキストデータの単語・トークンを割り当てる初期化クラス設定部と、前記単語・トークンの第一次元列から互いに異なる単語を抽出し、所定の出現頻度を有する単語のそれぞれに固有の単語クラスと、前記単語・トークンの第一次元列から互いに異なる単語を割り当てる初期化クラス設定部と、前記単語・トークンの第一次元列から互いに異なる単語クラスを算出する平均相互情報量算出部と、前記単語クラスの集合のうち、前記平均相互情報量が最大である2つの単語クラスを本マージする本マージ部とを備えることを特徴とする請求項6に記載の単語・連語分類処理装置。

【請求項8】前記単語クラス取出手段は、前記テキストデータの単語クラスの一次元列から、隣接して存在する2つ以上の単語クラスを順次に取り出し单語クラス取出部と、前記単語クラス取出部により取り出した2つ以上の単語クラスの相互情報量を算出する相互情報量算出部と、前記相互情報量が所定のしきい値以上の2つ以上の単語クラスをクラスチェック部で結合する。

【請求項9】前記単語・トークン分類手段は、前記テキストデータの単語・トークンの第一次元列から互いに異なる単語を算出する単語クラスと、前記テキストデータの単語・トークンクラスと互いに異なるトークンとを抽出し、所定の出現頻度を有する単語とトークンとのそれぞれに固有の単語・トークンクラスを削除する初期化クラス設定部と、前記テキストデータの単語・トークンクラスの熱集から2つの単語・トークンクラスについての平均相互情報量を算出する平均相互情報量算出部と、前記単語・トークンクラスの集合のうち、前記平均相互情報量が最大である2つの単語・トークンクラスを本マージする本マージ部とを備えることを特徴とする請求項6に記載の単語・連語分類処理装置。

【請求項10】テキストデータから連語を抽出する連語抽出手段と、前記テキストデータに含まれる単語と連語とを一

【特許請求の範囲】複数の単語の一次元列としてのテキストデータから、互いに異なる個の単語を抽出し、前記個々の単語の集合をC個の単語クラスに分割した第1のクラスリラングを生成するステップと、前記第1のクラスリラングに基づいて生成された前記テキストデータの単語クラスの一次元列において、隣接する単語クラス間の粘着度が全て所定値以上の単語クラス列の集合を抽出するステップと、前記単語クラス列に固有のトークンを対応させ、前記単語クラス列に属する単語列を前記テキストデータから検索し、前記テキストデータの単語列を対応するトークンと置換することにより、前記テキストデータに記載の単語と連語について

格に分類して、単語と連語が混在するクラスを生成する単語・連語分類手段などを備えることを特徴とする単語・連語分類処理装置。

【請求項1】前記クラスは、前記クラスの平均相互情報量に基づいて生成されることを特徴とする単語・連語の単語・連語分類処理装置。

【請求項1-1】前記クラスは、前記クラスから高遅延に連語を抽出することが可能な連語抽出装置の単語・連語分類処理装置。

【請求項1-2】テキストデータに含まれる単語を分類して単語クラスを生成することを特徴とする単語・連語の単語・連語分類処理装置。

【請求項1-3】前記クラスは、前記クラスの一次元列を生成する前記単語クラスを生成手段と、前記テキストデータ

データから抽出する単語クラスの一次元列を生成手段と、前記テキストデータの単語クラスの一次元列に

おいて、隣接する単語クラス間の粘着度が全て所定値以上の単語クラス列を、前記テキストデータの単語クラスの一

次元列から抽出する単語クラス列を構成する個々の単語クラスから、前記テキストデータ

データに隣接して存在する個々の単語を別々に取り出しうて連語を抽出することを特徴とする連語

抽出装置。

【請求項1-4】前記単語クラスは、前記単語クラスの平均相互情報量に基づいて生成されることを特徴とする請求項

1-2に記載の連語抽出装置。

【請求項1-5】所定のテキストデータに含まれる単語と連語とを、単語と連語とが混在するクラスに分類して格納して

いる単語・連語辞書と、前記単語・連語辞書と所定の隠れマルコフモデルとを参照することにより、発音音声を音声認

識する音声認識手段などを備えることを特徴とする音声認識装置。

【請求項1-6】所定のテキストデータに含まれる単語と連語とを、単語と連語とが混在するクラスに分類して格納して

いる単語・連語辞書と、前記単語と前記原文に対する用例訳文とを対応させて格納している用例文集と、入力さ

れた原文の単語が属するクラスと同一のクラスに属する単語又は連語により構成される用例原文を前記用例文集か

ら検索する用例訳文と同一のクラスに属する単語を、人力された原文の単語に対する訳語

に置換することにより、前記入力された原文に対する訳文を生成する用例適用手段とを備えることを特徴とする機械

翻訳装置。

【請求項1-7】テキストデータの単語の一次元列から互いに異なる単語と連語とを、前記クラスの平均相互情報量に基づいて生成されていることと特徴とする単語・連語記憶媒体。

【請求項1-8】テキストデータの単語の一次元列を抽出し、抽出された前記単語の集合を分割して単語クラスを生成する機能と、前記テキストデータの単語の一次元列を構成する個々の単語を前記単語が属する前記単語クラスで置換することにより、前記テキストデータの単語クラスの一次元列を生成する機能と、前記テキ

ストデータの単語クラスの一次元列から、隣接する単語クラス間の粘着度が全て所定値以上の単語クラス列を抽出す

る機能と、前記単語クラス列にトーケンを付与する機能と、前記単語の一次元列を生成する機能と、前記テキ

ストデータの単語クラス列を前記トーケンで置換することにより、前記テキストデータの単語・トーケンの一次元列を生成する機能と、前記テキストデータの単語・トーケンとが混在する集合を分割して単語・トーケンクラスを生成する機能と、前記テキストデータの単語・トーケンクラスの中のトーケンを、前記テキストデータに隣接して置換して連語を生成する機能と、前記コンピュータ読み取り可能な記憶媒体。

(0002)

【発明の詳細な説明】

(0001)

【発明の技術】從来の単語分類処理装置には、例えば、「Brown, P., Della Pietra, V., deSouza, P., Lai, J., Mercer, R. (1992) "Class-Based n-gram Models of Natural Language", Computational Linguistics, Vol. 18, No. 4, pp. 467-479」に記載されているように、テキストデータの中で使用されていいる単語の単語を統計的に処理することにより、単語と連語を自動的に分類するものがあり、この単語の単語認識装置や機械翻訳を行っていた。

(0003)

【発明が解決しようとする課題】しかしながら、從来の単語分類処理装置は、単語と連語とをまとめて自動的に分類することができず、単語と連語あるいは連語と連語を用いて、音声認識や機械翻訳を行うことかきないため、音声認識や機械翻訳を正確に実行することができないという問題があつた。

【0004】そこで、本発明の第1の目的は、単語と連語などをまとめて自動的に分類することができる単語・連語分類処理方法及び単語・連語分類処理装置。

【0005】また、本発明の第2の目的は、大量のテキストデータから高速に連語を抽出することができる連語抽出装置を提供することである。また、本発明の第3の目的は、単語と連語あるいは連語の対応関係や類似度を用いることにより、正確な音声認識が可能な音声認識装置を提供することである。

【0006】また、本発明の第4の目的は、単語と連語あるいは連語と連語の対応関係や類似度を用いることにより、正確な機械翻訳が可能な機械翻訳装置を提供することである。

(0007)

【課題を解決するための手段】上記した第1の目的を達成するために、本発明によれば、テキストデータに含まれる單語と連語とを一緒に分類して、単語と連語とが混在するクラスを生成するようにしている。

【0008】にこににより、単語と単語とをまとめて分類するだけでなく、単語と連語あるいは連語と連語とを一緒に分類することができ、単語と連語を容易に判別することができる。

【0009】また、本発明の一態様によれば、単語を分類した単語クラスをテキストデータの単語の一次元列にマッピングして単語クラスの一次元列を生成し、テキストデータの単語クラスの一次元列において、隣接する単語クラス間の粘着度が全て所定値以上の単語クラス列を抽出してその単語クラス列と付与し、単語とトーケンとと一緒に分類してから、トーケンに対応する単語クラス列をその単語クラス列に隣接するようにしている。

【0010】にこににより、単語クラス列にトーケンを付け与てその単語クラス列を1つの単語と連語との区別なく分類処理を行うに含まれる単語とトーケンとを取り扱って単語と連語との区別なく分類処理を行うことができる。また、単語を分類した単語クラスをテキストデータの単語の一次元列にマッピングして単語クラスの一次元列を生成し、隣接する単語クラス間の粘着度に基づいて連語を抽出することにより、テキストデータから抽出を高速に行うことができる。

【0011】また、上述した第2の目的を達成するために、本発明によれば、単語を分類した単語クラスをテキストデータの単語の一次元列にマッピングして単語クラスの一次元列を生成し、テキストデータの単語クラスの一次元列において、隣接する単語クラス間の粘着度が全て所定値以上の単語クラス列を抽出して連語を抽出するようしている。

【0012】にこににより、単語クラス列に基づいて連語を抽出する個々の単語を抽出し、単語クラス列を隣接する個々の単語クラスから、テキストデータに隣接して存在する異なる単語の単語クラス列を抽出して連語を抽出するようしている。

【0013】にこににより、単語クラスの数が少ないので、テキストデータの単語クラスの数のほうが少ないので、テキストデータの単語の一次元列において、隣接する単語クラス列を抽出するほうが、テキストデータの単語の一次元列に比べて、演算量及びメモリ容量を少くすることができますが、連語の抽出処理を高速に行うことができるとともに、メモリ資源を節約できる。なお、単語クラス列には、テキストデータの単語の一次元列に存在しない単語列が含まれている場合があるので、単語クラス列を構成する個々の単語クラスから、テキストデータに隣接して存在する個々の単語を別々に取り出して連語としている。

【0014】また、上述した第3の目的を達成するために、本発明によれば、所定のテキストデータに含まれる単語と連語とを、単語と連語を参照することにより、発音音声を音声認識するようにして、連語辞書を参照することにより、発音音声を音

声認識するようになっている。

【0015】にこににより、単語と連語あるいは連語の対応関係や類似度を用いながら音声認識を行なうことで、正確な処理が可能になる。また、上述した第4の目的を達成するために、本発明によれば、所定のテキストデータに含まれる単語と連語とを、単語と連語とが混在するクラスに分類して格納している単語・連語辞書に基づいて、用例文集に基づいて格納している単語・連語辞書とを对照して連語と連語とが混在するクラスに分類するものである。

【0016】図1は、本発明の一実施例に係わる単語・連語分類処理装置について図面を参照しながら説明する。図1において、単語分類手段1は、所定のテキストデータの単語の一次元列から異なる単語と連語とを、単語と連語とが混在するクラスに分類する集合を分割して単語クラスを生成する。

【0017】図2は、本発明の一実施例に係わる単語・連語分類処理装置のプロック図である。図1において、単語分類手段1は、用例文集に基づいて格納されている用例原文と入力された原文とを対応させることによって、用例文集に基づいて格納されている用例原文の単語が連語と連語とが混在するクラスに置き換わった原文が入力された場合においても、入力された原文に用例原文を適用して機械翻訳を行うことができる。

【0018】図1は、本発明の一実施例の処理を説明するもので、テキストデータに含まれる単語の単語と連語とが混在するクラスに並べたV個の単語がキャラクターとしての単語の一次元列(w1:w2 w3:w4 …:wT)から、テキストデータでの出現頻度順に並べたV個の単語の

で結ばれた単語クラス例は、例えば、C1 - C3、C1 - C7、…、C2 - C3、C2 - C11、…、C300 - C32、…、C1 - C3 - C80、C1 - C4 - C5、C3 - C7 - C2、…、C1 - C9 - C11 - C32、…とする。この場合、単語クラス例C1 - C3 に対してトークン1 を付与し、単語クラス例C1 - C7 に対してトークン2 を付与し、…、単語クラス例C2 - C3 に対してトークン3 を付与し、単語クラス例C2 - C11 に対してトークン4 を付与し、…、單語クラス例C300 - C32 に対してトークン5 を付与し、…、単語クラス例C1 - C3 - C80 に対してトークン6 を付与し、…、単語クラス例C1 - C4 - C5 に対してトークン7 を付与し、単語クラス例C3 - C7 - C2 に対してトークン18 を付与し、…、単語クラス例C1 - C9 - C11 - C32 に対してトークン19 を付与する。

[0042] 図1の単語トークン列生成手段5は、テキストデータの単語の一次元列(w1 w2 w3 w4 w5 w6 w7 … wT)のうち、単語クラス例抽出手段4により抽出された単語クラス列に属する単語列をトークンで置換することにより、テキストデータの単語トークンの一次元列を生成する。

〔例〕9(c)は、テキストデータの単語トークンの一次元列の一例を英文を例にとって示す図である。図9(b)のテキストデータの単語の一次元列(w1 w2 w3 w4 w5 w6 w7 w8 w9 w10 w11 w12 w13 w14 w15)として、図9(a)の "He wanted the apartment by bus and she went to New York by plane" が対応しているものとし、この単語の一次元列(w1 w2 w3 w4 w5 w6 w7 w8 w9 w10 w11 w12 w13 w14 w15)に1対1に対応する単語クラスの一次元列が図9(c)の(C5 C90C3 C21C18C101 C32C2 C5 C90C3 C63C28C101 C32)で与えられるものとする。

[0046]この單語クラスの一次元列(C5 C90C3 C21C8C101 C32C2 C5 C90C3 C63C28C101 C32)において、隣接する2つの單語クラス(C_i , C_j)が相互情報量MI(C_i , C_j)を計算し、相互情報量MI(C_i , C_j)が所定のしきい値TH以上、相互情報量MI(C₅ , C₉₀)、MI(C₉₀ , C₃)、MI(C₃ , C₂₁)、MI(C₂₁ , C₁₀₁)、MI(C₁₀₁ , C₃₂)、MI(C₃₂ , C₂)、MI(C₂ , C₅)、MI(C₅ , C₃₀)、MI(C₃₀ , C₃)、MI(C₃ , C₆₃)、MI(C₂₈ , C₁₀₁)及びMI(C₁₀₁ , C₃₂)が所定のしきい値THより小さい場合、隣接する2つの單語クラス(C₆₃ , C₂₈)が、図9 (d)に示すように、クラスチェーンで結ばれる。

w2, w3, w4, ..., wN 又は 1 個のトークンの単語 [t1, t2, t3, t4, ..., tD] を分割する。トーカンが混在して存在するD個の単語、トークンクラス [T1, T2, T3, T4, ..., TD] を生成する。

[0049] この単語トークン分類手段6では、トークンを付与された単語クラス列が1つの単語のようにみなされ、テキストデータに含まれる単語 [w1, w2, w3, w4, ..., wN] とトークン [t1, t2, t3, t4, ..., tL] とを同等に取扱う。

り扱うことができる。單語 $w_1, w_2, w_3, w_4, \dots, w_N$ とトーケン [1], [2], [3], [4], \dots, [L] との区別なく分類処理を行うことができる図10は、図1の単語-トーケン分類手段6の機能的な構成を示すブロック図である。

[0050] 図10において、初期クラス設定部40は、テキストデータの単語・トークン列から互いに異なる単語と互いに異なるトークンを抽出し、所定の出現頻度を有するN個の単語(w1、w2、w3、w4、…、wN)と1個のトークン(h1、t2、t3、t4、…、tl)とのそれぞれに固有の単語・トークンラス[T1、T2、T3、T4、…、Tl]を割り当てる。

[0051] 版マージ部41は、単語・トークンラスの集合[T1、T2、T3、T4、…、Tl]から2つの単語・トークン

[0052] 平均相互情報量算出部42は、テキストデータの仮マージされた単語・トークンクラス[T1、T2、T3、T4、…、TM-1]についての平均相互情報量AMIを(1)式により算出する。この場合、M個の単語クラス・トークンクラスの集合[T1、T2、T3、T4、…、TM]から、2つの単語・トークンクラス[Ti、Ti']を取り出す取り出したかたは、 $M(M-1)/2$ 個だけ存在するので、 $M(M-1)/2$ 回の平均相互情報量AMIの計算を行う必要がある。

[0053] 本マージ部43は、仮マージにより計算されたM(M-1)／2個の平均相互通報量AMIの基づいて、平均相互通報量AMIを最大とする2つの単語・トークンクラス[T1, Tj]を単語クラス・トークンクラスの集合[T1, T2, T3, T4, ..., TM]から取り出して本マージする。このことにより、本マージされたいずれかの単語・トークンクラス[T1, Tj]に属する単語及びトークンは、同一の単語クラス・トークンクラスに分類される。

[0054] 図1の連語置換手段7は、単語・トークン列生成手段5により置換された単語列に逆置換して連語を生成する。図11は、クラスチェーンと連語との関係を説明する図である。

[0055] 図1において、例えば、単語クラスC300と単語クラスC32どがクラスチェーンで結ばれ、このクラスチェーンで結ばれた単語クラスC300とC32にトークン15が付されているとする。また、単語"Toyota"、"Nissen"、"GM"…などのA個の単語が単語クラスC300に属し、単語"car"、"track"、"wagon"…などのB個の単語が単語クラスC32に属しているとする。

[0056]にの場合、連語の候補として、図11(b)に示すように、“Toyota car”、“Toyota track”、“Toyota wagon”、“Nissan car”、“Nissan track”、“Nissan wagon”、“GM car”、“GM track”、“GM wagon”、“など、単語クラスC300に属するA個の単語と単語クラスC32に属するB個の単語との順列の組合せで候補が生成される。この連語の候補の中にテキストデータに存在しない連語も含まれているので、テキストデータをスキャンすることで、これら連語の候補からテキストデータに存在する連語のみを抽出する。例えば、テキストデータには、“Nissan track”及び“Toyota wagon”は存在するが、“Toyota car”、“Toyota track”、“Nissan car”、“Nissan an wagon”、“GM car”、“GM track”及び“GM wagon”は存在しない場合、図11(c)に示すように、“Nissan track”及び“Toyota wagon”的みが連語としてテキストデータから抽出される。

[0057] 図12は、C1の単語クラス[C1, C2, C3, C4, …, CC]、Dの単語・トークンクラス[T1, T2, T3, T4, …, TD]及びDの単語・単語クラス[R1, R2, R3, R4, …, RD]の一例を示す図である。

[0058] 図12(a)において、C1の単語クラス[C1, C2, C3, C4, …, CC]が、図1の単語分類手段11により生成され、例えば、“he”, “she”, “it”などの単語が単語クラスC5に属し、“York”, “London”…などの単語が単語クラスC28に属し、“new”, “track”, “wagon”…などの単語が単語クラスC32に属し、“new”, “old”…などの単語が単語クラスC69に属し、“Toyota”, “Nissan”…などの単語が単語クラスC800に属する。このように、単語データベース上に記述された各単語が、各単語クラスに属する。

【0060】トークン付手手段4は、単語クラス別C63-C28にトークン1を付与し、単語クラス別C300-C321にトークン2を付与する。

トドを切り下す。単語・トークン生成手段10は、テキストデータの単語の一次元列(w1 w2 w3 w4 ...wn)に存する“New York”と“track”を“Toyota wagon”で置き換える。テキストデータの単語の一次元列を生成する。“London”、“London”、“she”、“it”、“[0061]”などの単語及び“t1”、“t5”などのトークンについての分類処理を行い、図12(b)の0回ar、“track”、“wagon”…などの単語及び“t1”、“t5”などのトークンの一次元列に存在する“he”、“she”、“it”、“London”、“London”を抽出する。

[0062]単語・トークンクラス[T1, T2, T3, T4, ..., TD]において、例えば、“he”、“she”、“it”...などの単語やトークンが単語トークンクラスT5 に属し、“I”, “London”...などの単語やトークンが単語・トークンクラスT28 の単語トークンクラス[T1, T2, T3, T4, ..., TD]を生成する。

に属し、“car”、“track”、“wagon”、“t5”…などの単語やトークンが単語トークンクラスT32に属し、“new”、“old”…などの単語やトークンが単語トークンクラスT63に属し、“Toyota”、“Nissan”、“GM”…などの単語やトークンが単語トークンクラスT300に属している。このように、単語トークンクラス[T1, T2, T3, T4, …, TD]には単語とトークンとの区別なく、単語とトークンが混在して分類されている。

5 “などのトークンを、テキストデータの単語の一次元列に存在する連語で置換することにより、図12(c)の單語-連語クラス[R1, R2, R3, R4, …, RD]に属しているトークン[1]は、単語-トークン列生成手段5により、テキストデータの単語の一次元列に存在する “New York” を置換されたものなので、このトークン[1]を “New York” で置換することにより、単語-連語クラス[R2, R3]を生成し、单語-トークン[2]～[4]に属するトークン[2]～[4]は、出区間-トークン[2]～[4]生成手段5により、テキストデータの単語の一次元列に存在する連語で置換することにより、図12(c)の單語-連語クラス[R1, R2, R3, R4, …, RD]に属しているトークン[2]～[4]を生成する。例えば、单語-トークン[1]～[4]を生成する。”

る "Nissan track" 及び "Toyota wagon" と置換されたものなので、このトーン16を "Nissan track" 及び "Toyota wagon" で置換することにより、単語・連語分類処理装置を実現するシステム構成を示すロック図である。図13において、単語・連語分類処理部41のメモリインターフェース42、46、CPU43、ROM44、ワークRAM45、RAM47、ドライバ71及び通信インターフェース72はバス48を介して互いに接続され、テキストデータ44が単語・連語分類処理部41に入力されると、ROM44に格納されているプログラムに従って、CPU43はテキストデータ40を処理し、テキストデータ40の単語及び連語の分類処理を行う。テキストデータ40の単語及び連語の分類処理結果は、単語・連語辞書49に格納される。なお、テキストデータ40や単語及び連語の分類処理結果を通信インターフェース72から通信ネットワーク73を介して送信したり、受信したりすることも可能である。

[0065]また、単語及び連語の分類処理を行なうプログラムを、ハードディスク74、ICメモリカード75、磁気テープ76、フロッピーディスク77またはCD-ROMやDVD-ROMなどの光ディスク78による記憶媒体からRAM47にロードした後、このプログラムをCPU43で実行させるようにしてほしい。

[0066]さらに、単語及び連語の分類処理を行なうプログラムを、通信インターフェース72を介して通信ネットワーク73から取り出しこともできる。通信インターフェース72と接続される通信ネットワーク73として、例えば、LAN(LocalArea Network)、WAN(Wide Area Network)、インターネット、アノログ電話網、デジタル電話網(ISDN: Integral Service Digital Network)、PHS(ハンディシッテム)や衛星通信などの無線通信網などを用いることが可能である。

[0067]図14は、図1の単語・連語分類処理装置の動作を示すフローチャートである。図14において、まず、ステップS1に示すように、単語クラスタリング処理を行う。この単語クラスタリング処理では、複数の単語の一次元列(w1 w2 w3 w4 ... wT)としてのテキストデータから、互いに異なるV個の単語[v1、v2、v3、v4、...、vV]を抽出し、V個の単語の集合[v1、v2、v3、v4、...、vV]をC個の単語クラス[C1、C2、C3、C4、...、CC]に分割する第1のクラスタリング処理を行う。

[0068]ここで、V個の単語[v1、v2、v3、v4、...、vV]それぞれに単語クラス[C1、C2、C3、C4、...、CC]を割り当ててから、V個の単語クラス[C1、C2、C3、C4、...、CC]についてマージ処理を行うことにより、V個の単語クラス[C1、C2、C3、C4、...、CC]の個数を1つずつ減らしてC個の単語クラス[C1、C2、C3、C4、...、CC]を生成する場合、Vが7000もの数とななるときは、マージ処理を行なう(1)式の平均相互情報量AMIの計算回数が莫大なものとなり、現実的ではなくなる。このため、ワンドウ処理を行なう。

[0069]図15は、ワンドウ処理を説明する図である。図15(a)において、テキストデータのV個の単語[v1、v2、v3、v4、...、vV]それぞれに割り当てられたV個の単語クラス[C1、C2、C3、C4、...、CC]のうち、テキストデータでの出現頻度の大きい単語に割り当てられたC+1個の単語クラス[C1、C2、C3、C4、...、CC、CC+1]についてのマージ処理を行う。

[0070]ここで、図15(b)に示すように、M個の単語クラス[C1、C2、C3、C4、...、CM]は、ワンドウ内のC+1個の単語クラス[C1、C2、C3、C4、...、CC、CC+1]についてのマージ処理を行なった場合、M個の単語クラス[C1、C2、C3、C4、...、CM]の数が1つ減ってM-1個の単語クラス[C1、C2、C3、C4、...、CM-1]となるとともに、ワンドウ内のC+1個の単語クラス[C1、C2、C3、C4、...、CC、CC+1]の数も1つ減ってC個の単語クラス[C1、C2、C3、C4、...、CC]となる。

[0071]この場合、図15(c)に示すように、ワンドウ外の単語クラス[CC+1、...、CM-1]のうち、テキストデータでの出現頻度が最も大きい単語クラス[CC+1]をワンドウ内に入れ、ワンドウ内の単語クラスの数が一定に保たれるようになる。

[0072]そして、ワンドウ外に単語クラスがなくなり、図15(d)のC個の単語クラス[C1、C2、C3、C4、...、CC]が生成された時に、単語クラスタリング処理を終了する。

[0073]なお、上述した実施例では、ワンドウ内の単語クラスの個数をC+1個に設定したが、C+1個以外のV個未満の数でもよく、また、途中で変化させようにもしてもよい。

[0074]図16は、ステップS1の単語クラスタリング処理を示すフローチャートである。図16において、まず、ステップS10に示すように、V個の単語の一次元列(w1 w2 w3 w4 ... wT)としてのテキストデータに基づいて、重複を除いた全てのV個の単語[v1、v2、v3、v4、...、vV]の出現頻度を調べ、これらのV個の単語[v1、v2、v3、v4、...、vV]を出現頻度の高い単語から順に並べて、これらのV個の単語[v1、v2、v3、v4、...、vV]のそれをV個の単語クラス[C1、C2、C3、C4、...、CV]に割り当てる。

[0075]次に、ステップS11に示すように、V個の単語クラス[C1、C2、C3、C4、...、CV]の単語のうち、出現

頻度の高い単語クラスの単語から、V個未満のC+1個の単語クラスの単語を1つのワンドウ内の単語クラスの單語とする。

[0076]次に、ステップS12に示すように、1つのワンドウ内の単語クラスの単語の中で、全ての組み合わせの版ペアを作り、各版ペアを版マージした時の平均相互情報量AMIを(1)式により計算する。

[0077]次に、ステップS13に示すように、全ての組み合わせの版ペアについての平均相互情報量AMIのうち、最大となる平均相互情報量AMIを有する版ペアを本マージし、本マージ後の1つのワンドウ内の単語クラスはC個になつたかどうかを判断し、この条件が成立しない場合、ステップS15に進み、現在のワンドウよりも外側にあり、最大の出現頻度を有するクラスの単語をワンドウ内に入れ、ステップS12に戻り、以上の処理を繰り返すことにより、単語クラスの数を減少させる。

[0078]次に、ステップS14に示すように、ワンドウ内の単語クラスはC個になつた場合、ワンドウ内の単語クラスはC個になつたかどうかを判断し、この条件が成立しない場合、ステップS15に進み、現在のワンドウよりも外側にあり、最大の出現頻度を有するクラスの単語をワンドウ内に入れ、ステップS12に戻り、以上の処理を繰り返すことにより、単語クラスの数を更新する。

[0079]一方、ステップS14の条件が成立し、ワンドウ外に単語クラスがなくなり、かつ、ワンドウ内の単語クラスはC個になつた場合、ワンドウ内のC個の単語クラス[C1、C2、C3、C4、...、CC]をメモリに記憶する。

[0080]次に、図14のステップS2に示すように、クラスチーン抽出処理を行う。このクラスチーン抽出処理では、ステップS1の第1のクラスチーンング処理に基づいて生成されたテキストデータの単語クラスの一次元列において、所定のしきい値以上の相互情報量を有する隣接する2つの単語クラスをチーンで結ぶにより、チーンで結ばれた単語クラス列の集合を世出す。

[0081]図17は、ステップS2のクラスチーン抽出処理の第1実施例を示すフローチャートである。図17において、まず、ステップS20に示すように、テキストデータの単語クラスの一次元列から、互いに隣接する2つの単語クラス[C1、C2、C3、C4、...、CC]を取り出す。

[0082]次に、ステップS21に示すように、ステップS20で取り出した2つの単語クラス(C1、C2、C3)についての相互情報量MI(C1、C2、C3)を(2)式により計算する。

[0083]次に、ステップS22に示すように、ステップS21で計算した相互情報量MI(C1、C2、C3)が所定のしきい値TH以上である場合、ステップS23に進んで、相互情報量MI(C1、C2、C3)が所定のしきい値TH以上である場合、ステップS23に進んで、相互情報量MI(C1、C2、C3)をクラスチーンで結んでメモリに格納し、相互情報量MI(C1、C2、C3)を取り出す。

[0084]次に、ステップS24に示すように、メモリに格納されているクラスチーンで結ばれた単語クラスにおいて、单語クラスCiで終了しているクラスチーンが存在するかどうかを判断し、单語クラスCiで終了しているクラスチーンが存在する場合、クラスチーンをつなぐ。

[0085]一方、ステップS24において、单語クラスCiで終了しているクラスチーンが存在しない場合、ステップS25。をスキップする。次に、ステップS26に示すように、テキストデータの単語クラスの一次元列から、互いに隣接する2つの単語クラス(C1、C2、C3)を全て取り出したかどうかを判断し、互いに隣接する2つの単語クラス(C1、C2、C3)を全て取り出して、单語クラスCiで終了しているクラスチーンが存在するかどうかを判断し、单語クラスCiで終了していない場合、ステップS25に進んで、单語クラスCiで終了しているクラスチーンをつなぐ。

[0086]図18は、ステップS22において、单語クラスCiで終了しているクラスチーンが存在しない場合、ステップS20に示すように、テキストデータの単語クラスの一次元列から、互いに隣接する2つの単語クラス(C1、C2、C3)を取り出し、このC+1個の単語クラス(C1、C2、C3、C4、...、CC、CC+1)についてのマージ処理を行なう。

[0087]次に、ステップS20に示すように、長さ2の全てのクラスチーンをそれぞれオブジェクトで置き換える。ここで、オブジェクトは、上述したトークンと同じものを表しているが、長さ2のクラスチーンを、特に、オブジェクトと呼ぶ。

[0088]次に、ステップS20に示すように、テキストデータのクラスチーンをオブジェクトで置き換え、テキストデータのクラスとオブジェクトの第一次元列を生成する。すなまち、テキストデータのクラスとオブジェクトのクラスとオブジェクトの第一次元列においての相互情報量MI(C1、C2、C3)は、互いに隣接する1つのクラスと1つのオブジェクト(長さ2のクラスチーン)との間で算出される場合、互いに隣接する1つのオブジェクト(長さ2のクラスチーン)との間で算出される場合、及び互いに隣接する1つのオブジェクト(長さ2のクラスチーン)と1つ

のオブジェクト長さ2のクラスチェーン)との間で算出される場合がある。

[0090]次に、ステップS205に示すように、ステップS204で計算した相互情報量MI(C_i、C_j)が所定のしきい値H以上であるかどうかを判断し、相互情報量MI(C_i、C_j)が所定のしきい値TH以上である場合、ステップS26に進んで、ステップS204で取り出した互いに隣接する2つのクラス、又は互いに隣接する1つのクラスと1つのオブジェクト、又は互いに隣接する2つのオブジェクトをクラスチェーンで結び、相互情報量MI(C_i、C_j)が所定のしきい値THより大きい場合は、ステップS206をスキップする。

[0091]図19は、テキストデータの一次元列におけるクラスチェーンを示す図である。図19において、互いに隣接する1つのクラスとオブジェクトの一次元列のクラスとオブジェクトの一次元列の構成を示す。図19のうち、単語列(w₁ w₂)、(w₁₃w₁₄)、…がトークン1に置換され、单語列(w₄ w₅ w₆)、(w₁₇w₁₈)、…がトークン2で置換されたとすると、トークン1に対応する連語として、(w₁ - w₂、w₁₃-w₁₄、…)-がテキストデータから抽出され、トークン2に対応する連語として、(w₄ - w₅ - w₆、w₁₇-w₁₈、…)-がテキストデータから抽出される。

[0092]次に、図18のステップS207に示すように、クラスチェーン抽出処理が所定の回数行われたかどうかを判断し、所定の回数行われていない場合は、ステップS202に戻って以上の処理を繰り返す。

[0093]このように、長さ2のクラスチェーンをオブジェクトに置き換えて、相互情報量MI(C_i、C_j)を算出することを繰り返すことにより、任意の長さのクラスチェーンを抽出することができる。

[0094]次に、図14のステップS31に示すように、トークン置換処理を行ふ。このトークン置換処理では、ステップS20のクラスチェーン抽出処理で抽出された單語クラス列に固有のトークンを対応させ、この單語クラス列に属する單語列をテキストデータの單語の一次元列から検索し、テキストデータの單語列を対応するトークンで置換することにより、テキストデータについての單語ヒートマップの一次元列を生成する。

[0095]図20は、ステップS33のトークン置換処理を示すフローチャートである。図20において、まず、ステップS30に示すように、抽出されたクラスチェーンを重複を除いて所定の規則でソートし、それぞのクラスチェーンにトークンを対応させて、クラスチェーンに名前を付ける。ここで、クラスチェーンのソートは、例えば、ASCIIコード順で行う。

[0096]次に、ステップS31に示すように、トークンに応応させたクラスチェーンを1つ取り出す。次に、ステップS32に示すように、テキストデータの單語の一次元列の中にクラスチェーンに属する單語クラス列に属する单語列が存在するかどうかを判断し、クラスチェーンに属する单語列を1つのトークンで置き換える。ステップS33に進み、テキストデータの対応する单語列を1つのトークンで置き換える、クラスチェーンで結ばれた单語クラス列に属する单語列が单語ヒートマップの一次元列の中に存在しなくなるまで以上の処理を繰り返す。

[0097]一方、クラスチェーンで結ばれた单語クラス列に属する单語列が存在しない場合、ステップS34に進み、ステップS30でトークンに対応させた全てのクラスチェーンについての連語トークン置換処理が終了したかどうかを判断し、全てのクラスチェーンについての連語トークン置換処理が終了しない場合、新たにクラスチェーンを1つ取り出して、以上の処理を繰り返す。

[0098]次に、図14のステップS4に示すように、单語トークンクラスタリング処理を行う。この单語トークンクラスタリング処理では、テキストデータの单語の一次元列において、互いに異なる单語と互いに異なるトークンとを抽出し、单語とトークンが混在する集合を单語トークンクラスタに分類する。

[0099]図21は、ステップS4の单語・トークンクラスタリング処理を示すフローチャートである。図21において、ステップS40に示すように、ステップS3で得られたテキストデータの单語トークンの一次元列を入力データとして、ステップS1の第1の单語クラスタリング処理と同一の方法でクラスリスト[T1、T2、T3、T4、…、TD]を生成する。この第2のクラスリスト[T1、T2、T3、T4、…、TD]を生成する。この第2のクラスリスト[T1、T2、T3、T4、…、TD]に分割する第3のクラスリスト[T1、T2、T3、T4、…、TD]は、その要素として单語とトークンを含んでいる。

[0100]次に、図14のステップS5に示すように、データ出力処理を行う。このデータ出力処理では、テキストデータの单語の一次元列に存在する单語列のうち、トークンに応応するものを連語として抽出し、单語・トークンクラス[T1、T2、T3、T4、…、TD]の中のトークンを連語で置換することにより、单語と連語とが混在する集合を单語・連語クラス[R1、R2、R3、R4、…、RD]に分割する。

[0101]図22は、ステップS5のデータ出力処理を示すフローチャートである。図22において、まず、ステップS50に示すように、1つの单語・トークンクラスTiから1つのトークンIKを取り出す。

[0102]次に、ステップS51に示すように、テキストデータの单語の一次元列をスキヤンシ、ステップS52において、ステップS50で取り出したトークンIKに対応するクラスチェーンで結ばれた单語クラス列に属する单語列が存在するかどうかを判断する。そして、トークンIKに対する用例文

ストデータの单語の一元列に存在する場合、ステップS53に進んで、この单語列を连語とみなす処理を繰り返し、テキストデータの单語の一次元列をスキャンすることにより得られた单語列をトークンIKを置き換える。

[0103]一方、トークンIKに対応するクラスチェーンで結ばれた单語クラス列に属する单語列がテキストデータの单語の一元列に存在しない場合、ステップS54に進んで、全てのトークンについて処理が終了したかどうかを判断し、全てのトークンについて処理が終了していない場合、ステップS50に進んで、以上の処理を繰り返す。

[0104]例えば、ステップS3のトークン置換処理において、テキストデータの单語の一次元列(w₁ w₂ w₃ w₄ … w_T)のうち、单語列(w₁ w₂)、(w₁₃w₁₄)、…がトークン1に置換され、单語列(w₄ w₅ w₆)、(w₁₇w₁₈)、…がトークン2で置換されたとすると、トークン1に対応する連語として、(w₁ - w₂、w₁₃-w₁₄、…)-がテキストデータから抽出され、トークン2に対応する連語として、(w₄ - w₅ - w₆、w₁₇-w₁₈、…)-がテキストデータから抽出される。

[0105]1つの单語・トークンクラスTiが单語の集合Wiとトークンの集合Ji = {i1, i2, …, itn}からなり、トークンクラスTiが{wi i1, i2, …, itn}により表され、トークンの集合Wiの中の1つの单語・連語クラスRiは、[0106]【数2】

[0107]で与えられる。以上説明したように、本発明の一実施例による单語・連語分類処理装置によれば、单語と連語を区別することができること。

[0108]次に、本発明の一実施例による音声認識装置について説明する。図23は、図1の单語・連語分類処理装置により得られた单語・連語分類処理部40に含まれる单語と連語とが、单語・連語辞書49に格納されている。

[0109]図23において、所定のテキストデータ40に含まれる单語と連語とが、单語・連語辞書49に格納された後、[0110]一方、複数の单語と連語とからなる发音聲音は、マイクロフォン50によりアナログ音声信号により単語と連語が混在するクラスに分類され、この分類された单語と連語とが单語・連語辞書49に格納されている。

[0110]A/D変換器51でデジタル音声信号に変換され、特徴抽出部52に投入される。特徴抽出部52は、デジタル音声信号に対して、例えば、LPC分析を行い、ケプストラム係数や対数ハーモニアなどの特徴パラメータを抽出する。特徴抽出部52で抽出された特徴パラメータは、音声認識部54に出力され、音素隠れマルコフモデル55を参照するとともに、单語・連語辞書49に格納されている单語と連語との分類結果を参照しながら、单語及び連語とに音声認識を行う。

[0111]図24は、单語・連語分類処理結果を示す図である。図24において、「本日は晴天なり」と発声された発音音声がマイクロフォン50に入力され、この発音音声に対して音声モデルを適用するとにより、例えば、「本日は晴天なり」という認識結果比(本日は晴天なり)という認識結果とが得られる。これらの音声モデルによる認識結果に対し、音語モデルによる処理を行つて单語・連語辞書49の参照を行い、「晴天なり」という連語が单語・連語辞書49に登録されている場合、「本日は晴天なり」という認識結果に対しても高い確率が与えられ、「本日は晴天なり」という認識結果に対する低確率が与えられる。

[0112]以上説明したように、本発明の一実施例による音声認識装置によれば、单語・連語辞書49を参照して音声認識を行うことにより、より正確な認識処理が可能になる。

[0113]次に、本発明の一実施例による機械翻訳装置について説明する。図25は、図1の单語・連語分類処理装置により得られた单語・連語分類処理結果を用いて機械翻訳装置の構成を示すブロック図である。

[0114]図25において、所定のテキストデータ40に含まれる单語と連語とが、单語・連語辞書49に格納された原文に対する用例原文に対する用例訳文とが、それぞれ対応させられて用例文集60に格納されている。また、用例原文とが混在するクラスに分類され、この分類された单語と連語とが单語・連語辞書49に格納されている。

[0115]用例文集部61に原文が入力されると、单語・連語辞書49を参照しながら入力された原文の单語が属するクラスを検索し、そのクラスと同一のクラスに属する单語又は連語はにより構成される用例原文を用例文集60から検索する。用例文集60から検索された用例原文及びその用例訳文の中の訳語を、入力された原文の单語に対する訳語に置換することにより、入力された原文に対する訳文が生成する。

[0116]図26は、单語・連語分類処理結果を利用して音声認識を行う場合の例を示す図である。図26において、「Toyota」と「Kohlberg Kravis Robert & Co.」とは同一のクラスに属し、「gained」と「lost」とは同一のクラスに属し、「2」と「1」とは同一のクラスに属しているものとする。

[0117]原文として、「Toyota gained 2 to 30 1/4」が入力されると、用例原文として、用例文集60から「Kohlberg Kravis Robert & Co. lost 1 to 80 1/2」が後述するとともに、その用例原文に対する用例文 「Kohlberg Kravis Robert & Co.」社は、1ドル値を下げる結果80 1/2ドルだった。」も検索される。

[0118]次に、用例原文の原語「Kohlberg Kravis Robert & Co.」と同一のクラスに属している入力原文の原語「Toyota」に対する訳語「トヨタ」で、用例原文の訳語「Kohlberg Kravis Robert & Co.」を置き換えることにより、入力原文に属する入力原文の原語「トヨタ」を「上げて」で、用例訳文の訳語「下げて」を置き換えることにより、「下げて」を「上げて」で置き換えて、用例訳文の数値「1」を「2」で置き換えることにより、「下げて」を「上げて」で置き換える。用例訳文の数値「80 1／2」を「30 1／4」で置き換えることにより、「下げて」を「上げて」で置き換える。

[0119]以上説明したように、本発明原文に対する訳文[トヨタは、2ドルを上げて終値30 1／2ドルだった。]を出力する。

[0120]以上、本発明の一実施例について説明したが、本発明は上述した実施例に限られるものではなく、本発明の技術的思想の範囲内で他の様々な変更が可能である。例えば、上述した実施例では、単語・連語分類処理装置を音声認識装置及び機械翻訳装置に適用した場合について説明したが、単語・連語分類処理装置を文字認識装置に用いるようにしてもよい。また、上述した実施例では、単語と連語とを混在される分類する場合について説明したが、連語のみを抽出し、この抽出した連語を分類するようにしてもよい。

[0121]〔発明の効果〕以上説明したように、本発明の単語・連語分類処理装置によれば、テキストデータに含まれる単語と連語とを一緒に分類して、単語と連語とが混在するクラスを生成することにより、単語と連語とをまとめて分類するだけでなく、単語と連語あるいは連語と連語とをまとめて分類することができ、単語と連語あるいは連語と連語との対応関係や類似度を容易に判別することができる。

[0122]また、本発明の一実施例によれば、テキストデータの単語クラス列にトークンを付与して単語クラス列を1つの単語とみなし、テキストデータに含まれる単語とトークンを付与された単語クラス列とを同等に取り扱ってこれらを分類しながら、テキストデータに存在する単語列で対応する単語クラス列を置き換えるようになりますので、単語と連語との区別なく分類処理を行うことができるとともに、テキストデータからの連語の抽出を高速に行なうことができる。

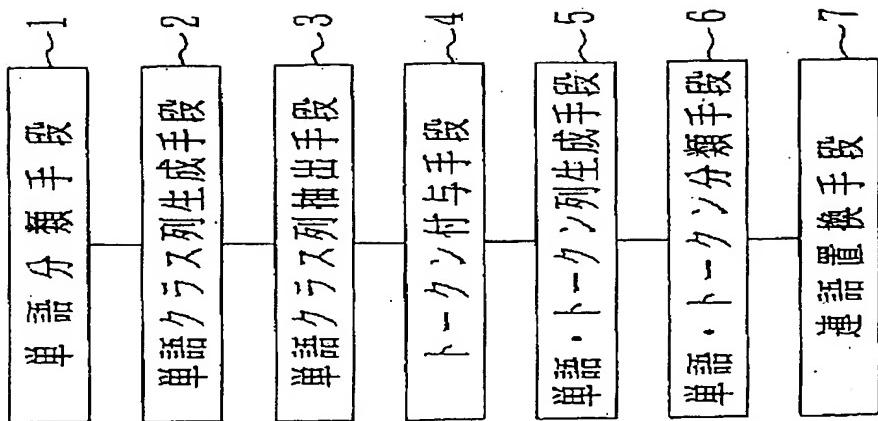
[0123]また、本発明の連語抽出装置によれば、テキストデータの単語列を構成する個々の単語を、その単語が属する単語クラスで置換し、テキストデータの単語クラス列を抽出してから、テキス

[0124]トデータに存在する連語を抽出することにより、連語を高速に抽出することができる。

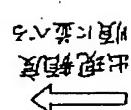
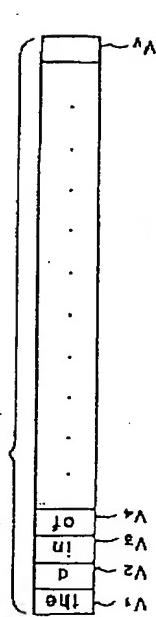
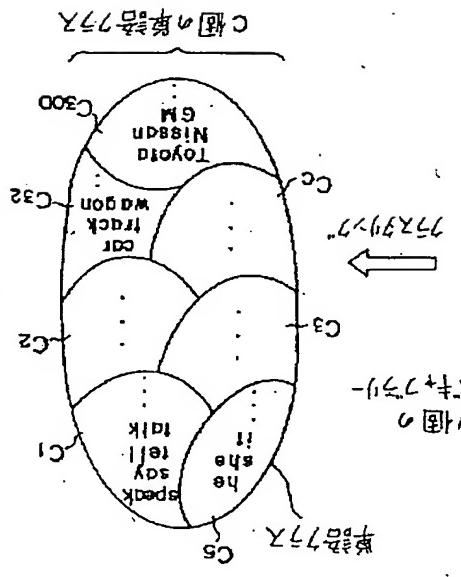
[0125]また、本発明の音声認識装置によれば、単語と連語あるいは連語と連語の対応関係や類似度を用いながら音声認識を行うことができる、正確な処理が可能になる。

[0126]また、本発明の機械翻訳装置によれば、用例文集に格納されている用例原文の単語が連語に置き換わった原文が入力された場合においても、入力された原文に用例原文を適用して機械翻訳を行うことができ、単語と連語あるいは連語と連語の対応関係や類似度を用いた正確な機械翻訳が可能になる。

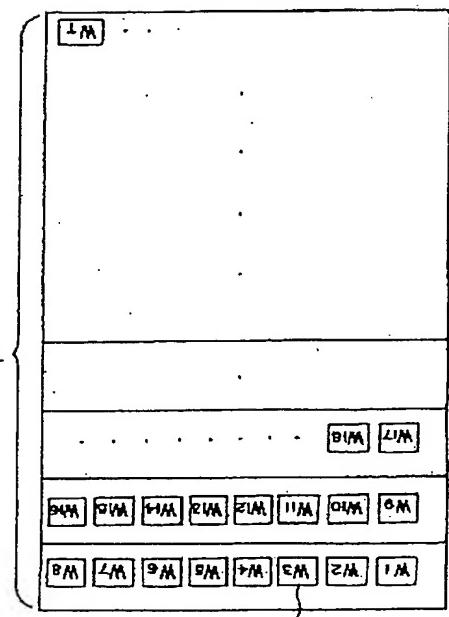
本発明の一実施例に係わる単語・連語分類処理装置の機能的な構成を示すブロック図



本発明の一実施例に係わる単語・連語分類処理装置の構成を図1に示す。



卷之三



四
卷之三